VISIT…

LANZAROTE Caliente.com

Материалы XII Всеросс. семинара «Нейроинформатика и её приложения», Красноярск, 2004. – 196с. – С.163-165.

РЕДУКЦИЯ РАЗМЕРОВ НЕЙРОСЕТИ НЕ ПРИВОДИТ К ПОВЫШЕНИЮ ОБОБЩАЮЩИХ СПОСОБНОСТЕЙ

Царегородцев В.Г.
www.NeuroPro.ru tsar@neuropro.ru

Для нейросетей обратного распространения ошибки при применении алгоритмов обучения, явно не направленных на максимизацию обобщающих способностей нейросети, редукция размеров нейросети (числа нейронов в скрытом слое, числа входных сигналов) чаще всего не приводит к росту обощающих способностей. Это противоречит мнению о том, что исключение шумовых, неинформативных признаков и избыточных нейронов является обязательным и полезным на практике.

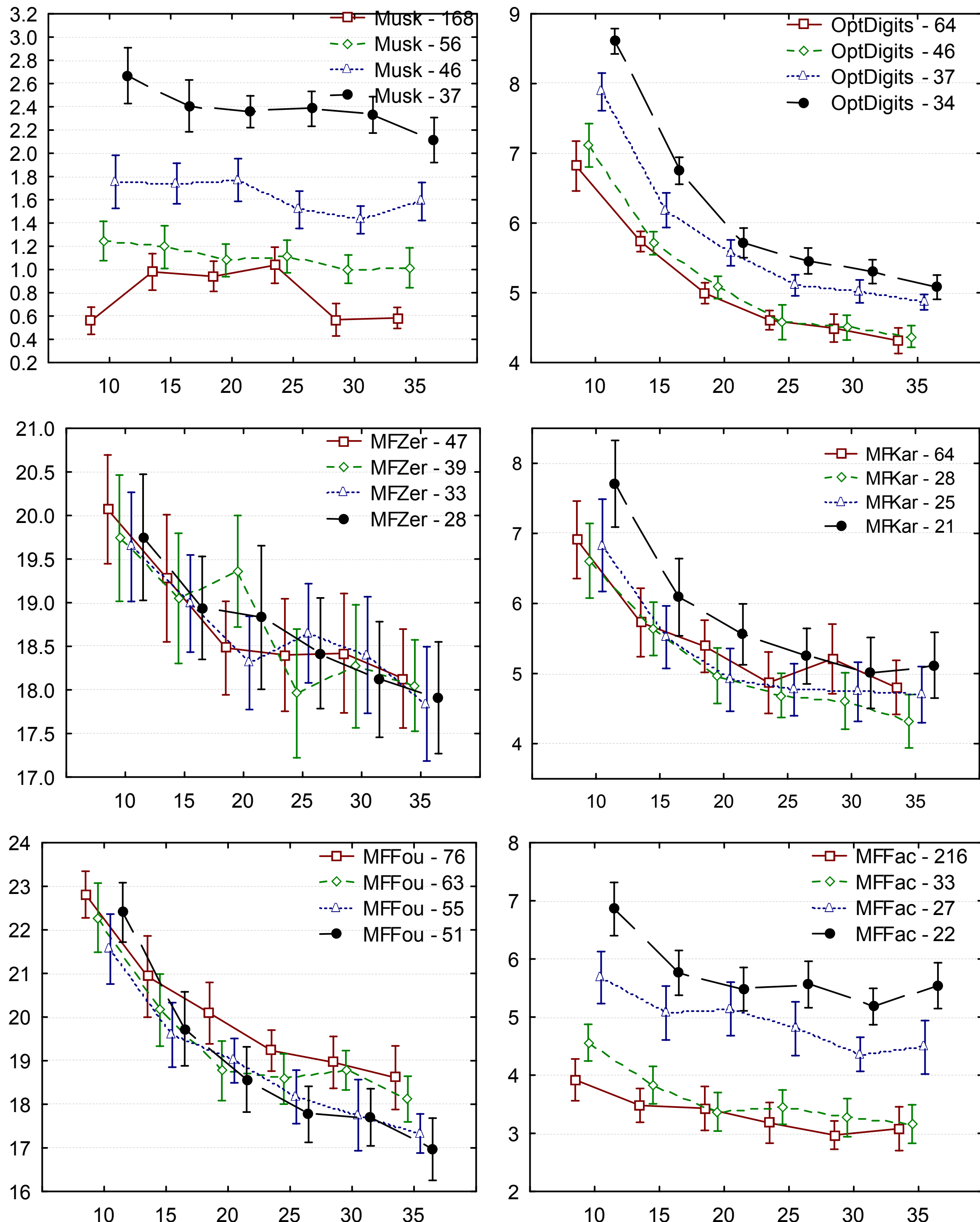
Постановка вопроса и результаты экспериментов

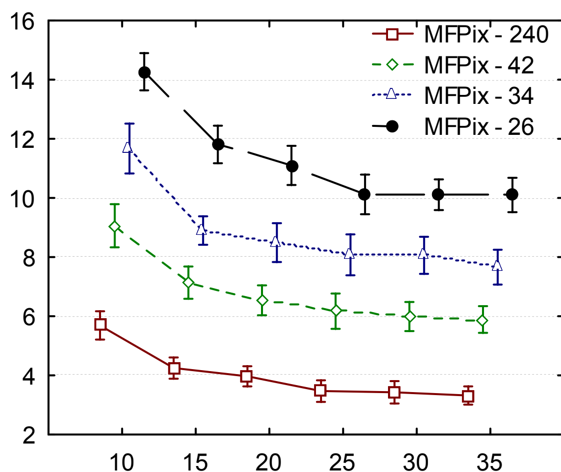
Работы [1,2] говорят, что причиной возникновения переобучения нейросети часто являются традиционные алгоритмы обучения, особенно наиболее “быстрые” из них, такие, как обучение по суммарному градиенту и особенно надстройка над последним метода наподобие сопряженных градиентов. Для борьбы с этим эффектом обычно используют уменьшение размеров нейросети до того момента, пока сохраняется требуемая точность решения обучающей выборки. В плане уменьшения размеров сети применяют редукцию числа независимых переменных, исключая неинформативные, избыточные и шумовые признаки, и редукцию лишнего числа нейронов в скрытых слоях, повышающих информационную ёмкость нейросети и поэтому могущих способствовать запоминанию шумов и выбросов в данных.

Здесь исследуется само базовое предположение – о снижении (улучшении) ошибки обобщения одновременно с уменьшением избыточности нейросети – уменьшением до того момента, когда ограничения размера станут препятствовать достижению нужного уровня ошибки обучения. Еще более специфически – работа инспирирована изучением [3] и целью своей ставит выдвижение и проверку вопроса о том, *соответствует ли минимум ошибки обобщения минимально необходимому для задачи числу признаков* (с возможным увеличением этого числа на некоторое малое число признаков), *либо минимум ошибки обобщения достигается вдали от минимально требуемого для меморизации задачи числа признаков*, т.е. при значительно бóльшем числе признаков. Если второе, то при редукции числа неинформативных переменных необходимо постоянно после очереного исключения малого числа переменных (одной или нескольких) проверять уровень ошибки обобщения и строить график наподобие представленного в [3], ну а иначе лишних промежуточных проверок делать не надо и можно автоматически элиминировать все избыточные для задачи признаки. Одновременно подобный вопрос ставится и для числа нейронов в скрытых слоях нейросети – нужно ли уменьшать число нейронов и как определять оптимальное число, особенно для MTL-парадигмы в свете замечания [4] о том, что для каждой из одновременно решаемых задач момент переобучения может наступить еще и в разное время (читай: через разное число эпох обучения).

У семи реальных задачах классификации из UCI KDD Database Repository (http://kdd.ics.uci.edu/) наличествует избыточное число независимых признаков, и поэтому можно проверить ошибки обобщения в нескольких достаточно сильно отстоящих друг от друга точках на оси числа признаков – для всего набора признаков, для минимального набора, для пары промежуточных значений. В дополнение к разным значениям признаков проверяется и несколько различных значений числа нейронов в скрытом слое сети – от 10 до 35 с шагом в 5 нейронов (10 нейронов достаточно для 100%-правильной классификации обучающей выборки в каждой задаче, т.е. размер

сети стартует с минимального размера или чуть завышенного). Были взяты базы Musk (168 независимых признаков), OptDigits (64 признака), 5 вариантов задачи Multiple features: Zer (47), Kar (64), Fou (76), Fac (216) и Pix (240 признаков). Далее на рисунках горизонтально отложено число нейронов, вертикально – ошибка обобщения (процент неправильных ответов на тестовой выборке), полученная при обучении сети на 4/5 базы данных и тестировании на оставшейся 1/5 части. Для каждой задачи, каждого размера сети и числа входов было обучено 25 нейросетей при различном разбиении базы данных на обучающую и тестовую выборки, по этим 25 результатам и рассчитаны средние значения ошибок обобщения и доверительные интервалы для них (доверительные интервалы отображаются в виде "усов" вокруг каждой точки графика). В строках легенд указаны числа признаков для каждого графика.





Два промежуточных (между минимальным и исходным значениями) числа признаков для каждой задачи выбирались по результатам действия алгоритма оценки информативности признаков – на графике информативности находилась пара переломных точек.

На всех семи графиках видно, что увеличение размера сети сверх реально достаточного для запоминания обучающей выборки размера в 10 нейронов приводит к росту обобщающей способности нейросети. Это же было отмечено и в [3].

Уменьшение ошибки обобщения с ростом размера сети наблюдается при любом числе входных сигналов. Таким образом, избыточный размер сети выступает в некотором роде как регуляризатор решения.

С числом независимых признаков ситуация немножко сложнее. Для одной задачи (Multiple features – Zer) никакого достоверного изменения ошибки обобщения ни в ту, ни в другую сторону не возникло. Для двух задач (Musk, Multiple features – Pix) даже число сигналов, превышающее на полтора-два десятка минимально необходимое число сигналов, ухудшает ошибку обобщения, и чем больше число входов сети приближается к минимальному числу, тем сильнее ухудшение обобщения. На паре задач (OptDigits, Multiple features – Fac) ухудшение обобщения начинает возникать при числе входов, превышающих минимальное число на десяток и менее. В задаче Multiple features – Kar все-таки наблюдается улучшение ошибки обобщения при “среднем”, т.е. избыточным по сравнению с минимальным, числе признаков, но только при избыточном размере сети. Близкие к минимальным числа признаков дают улучшение ошибки обобщения только в единственной задаче (Multiple features –Fou), но опять-таки только при увеличении размера сети.

Стабильность результатов при разных задачах (и разных объемах выборок в этих задачах) говорит, что при обучении нейросети традиционными алгоритмами, минимизирующими только ошибку аппроксимации обучающей выборки, но не эмпирическую ошибку, не нужно стремиться к уменьшению размера нейросети. Стремление к использованию числа признаков, близкого к минимально необходимому для аппроксимации требуемого отклика на обучающей выборке, также чаще всего может ухудшить эмпирическую ошибку.

Возможно, значительный вклад в формирование негативной зависимости ошибки обобщения от числа признаков дал алгоритм оценивания информативности признаков, оперирующий с нейросетью, т.е. несвободный от модели. Поэтому вдобавок желательно оценивать информативность только по выборке, до обучения нейросети.

Литература

1. *Fukumizu K.* Effect of batch learning in multilayer neural networks / Proc. 5th Int. Conf. Neural Information Processing (ICONIP’1998). 1998. – pp.67-70.
2. *Lawrence S., Giles C.L.* Overfitting and neural networks: conjugate gradient and backpropagation / Proc. Int. Joint Conf. Neural Networks (IJCNN'2000), Como, Italy. 2000. – pp.114-119.
3. *Caruana R.A., de Sa V.R.* Benefitting from the variables that variable selection discards / Journal of Machine Learning Research. 2003. Vol.3. – pp.1245-1264.
4. *Caruana R.A.* Multitask learning / Machine Learning. 1997. Vol.28. – pp.41-75.

Комментарии автора

Этот текст был написан в рамках борьбы на идеологическом и методологическом поле с местными оппонентами-конкурентами и предназначался для очного прочтения перед этими лицами, соответственно обязательно хочу предварительно отметить две вещи:

1. Изложение остановилось после выдвижения эмпирической гипотезы и получения экспериментальных результатов, теоретического объяснения я не дал, хотя прекрасно себе представлял. Просто нужно было выбить из рук оппонентов используемые ими методы и алгоритмы и запретить нажимать определенные «кнопки» в нейропрограммах, а объяснения результатов пусть они сами придумывают в меру своих сил, как и новые методы взамен дискредитированных старых.
2. Поскольку изложение было ориентировано на знакомую аудиторию, то не было сказано, что в работе была использована специализированная целевая функция (ЦФ) для задачи классификации (эта ЦФ отличается как менее жестким кодированием требуемых выходных сигналов сети по сравнению с жесткими кодами в МНК, так и реализацией раннего останова обучения). Широкие же массы используют другой софт и либо традиционный МНК для классификации и регрессии, либо кросс-энтропийную функцию для классификации. Результаты с такими целевыми функциями могут быть даже противоположными, хотя и для этих ЦФ дополнительными методами описанный эффект получить тоже можно.

Соответственно с п.2 приходится заметить, что это как раз сети наименьшего размера были «переучены» путем укладывания их в прокрустово ложе по требуемой точности (от сети требовалось обучение до момента 100%-правильного распознавания обучающей выборки), соответственно сети большего размера обучались именно до того же самого качественного уровня требований (куда уж лучше 100% распознавать?), но никак не всё более и более точному воспроизведению кодов классов, т.е. срабатывал вариант ранней остановки обучения. Тем не менее, на тестовой выборке результат сеток разного размера различался – что позволяет привязывать этот изменяющийся результат к свойствам, меняющимся при смене размера сети, а не к требуемой точности решения или алгоритму обучения (т.е. результаты в значительной степени всё же экстраполируемы и на другие условия).
А что же меняется с изменением размера сети? Да внутренние свойства (распределения свойств весов синапсов, производные), теоретическая зависимость ошибки обобщения от которых давно уже установлена [1,2]. Соответственно, на это в первую очередь нужно обращать внимание, а не на подгонку размера сети под требования к точности аппроксимации обучающей выборки.
Что же касается улучшения с ростом числа избыточных входных сигналов, то даже в случае скоррелированных избыточных входных сигналов шум в них чаще всего гораздо менее скоррелирован (обычно он уникален в каждом таком канале подачи информации) – соответственно усреднение сигналов с многих каналов подавляет шум, а не заставляет использовать быть может ложную корреляцию между сигналом+шум в одном единственном оставленном канале и требуемым выходом (ложная корреляция вряд ли даст хорошее качество последующего обобщения). А полностью шумовые сигналы нейросетью в нормальной ситуации никогда и не используются – она на них обращает внимание при завышенных требованиях к точности решения, и шум реально может пойти в дело только для того, чтобы в ту позу из Камасутры, в которую сеть ставит пользователь, сеть смогла-таки встать.
Соответственно, и при обучении избыточной по числу нейронов сети алгоритм обучения обычно не коррелирует выходы этих избыточных нейронов, и хоть они и вносят относительно малый вклад в результат решения сетью задачи, усреднение их сигналов выходным нейроном гасит транзитный шум через них со входа сети. Т.е. вклад лишних нейронов (они полностью лишними не бывают – после их удаления из обученной сети сетке чаще всего требуется дообучение, т.е. нейроны сетью задействовуются в некоторой мере – сетке проще размазать решение более-менее равномерно по всем нейронам, чем сформировать среди нейронов группу реальных пахарей-стахановцев и группу нейронов-лодырей) в решение хоть и мал получается, но дорог.
Это основные комментарии, объяснения результата и указания на связи.

1. Murata T., Yoshizawa S., Amari S. Learning curves, model selection and complexity of neural networks / Advances in Neural Information Processing Systems 5 (1992). Morgan Kaufmann, 1993. - pp.607-614.
2. Bartlett P.L. For valid generalization, the size of the weights is more important than the size of the network / Advances in Neural Information Processing Systems 9 (1996). MIT Press, 1997. - pp.134-140.